Data Visualization on Amazon Sales Data

Data Preprocessing Summary:

The initial dataset contained 1465 entries with 16 columns. The preprocessing steps involved:

1. Column Renaming: `discounted price` was renamed to `discount price` and `img_link` to `image link` for clarity.

2. Handling Missing Values:

* Two rows with missing values in `rating count` were identified and removed.
* One row with a missing value in `rating` was identified and removed.
* After these operations, the dataset has no remaining missing values.

3. Data Type Conversion:

`discount price` and `actual price` columns, which were objects containing currency symbols ('₹') and commas (','), were cleaned by removing these characters and then converted to `float` type.

* `discount percentage` was cleaned by removing the '%' symbol and converted to `float`.
* `rating` was processed to extract the numerical part (before any '|' character) and then converted to `float`.
* 'rating count' was cleaned by removing commas (',') and converted to `int` type.

4. Feature Engineering (Sub-categories Extraction): A new column named `subcategories` was created by splitting the `category` column by the '|' delimiter, providing a more granular view of product categories.

5. Outlier Removal: Outliers in numerical columns (`discount price`, `actual price`, `rating`, `rating count`) were identified and removed using the Interquartile Range (IQR) method (1.5 * IQR rule). This step significantly reduced the dataset size from 1463 to 1001 entries, ensuring that subsequent analyses are not disproportionately affected by extreme values.

Key Findings from Data Transformation:

1. Average Actual and Discount Prices by Main Category:

* Car Motorbike has the highest average actual price (~₹4000) and discount price (~₹2339).
* Home&Kitchen and Health&PersonalCare also show relatively high average prices.
* OfficeProducts and Toys&Games have the lowest average actual and discount prices.

2. Top-Rated Products: The top 10 products by rating (all rated 4.5) include items like milk frothers, vegetable cutters, and screen protectors, indicating strong customer satisfaction for these specific products.

Insights from Visualizations:

1. Product Categories Comparison:

Average Actual Price by Main Category: Car&Motorbike products are, on average, the most expensive, followed by Home&Kitchen and Health&PersonalCare. This suggests varying price points across different product sectors.

* Average Rating by Main Category: OfficeProducts and Toys&Games have the highest average ratings, indicating high customer satisfaction despite their lower average prices. This implies that high ratings are not necessarily tied to high prices, but rather to perceived value and quality.

* Average Discount Percentage by Main Category: HomeImprovement and Electronics categories offer the highest average discount percentages, which could be a strategy to attract customers in competitive markets. OfficeProducts and Toys&Games have much lower average discounts.

2. Price Distributions:

* Discounted Price Distribution: The distribution of discounted prices is right-skewed, with most products clustered at lower price points and a few products with significantly higher discounted prices.

* Actual Price Distribution: Similar to discounted prices, the actual price distribution is also right-skewed, with a majority of products being in the lower to mid-price range.

3. Rating Distribution: The rating distribution is left-skewed, indicating that most products receive high ratings (between 3.5 and 4.5). Very few products have ratings below 3.0 after outlier removal.

4. Rating Count Distribution: The distribution of rating counts is also right-skewed, indicating that many products have a moderate number of reviews, while a smaller number of popular products have a very high number of reviews.

5. Relationship between Actual Price and Rating (Scatter Plot): The scatter plot revealed no strong linear correlation between actual price and product rating. High ratings are observed across various price points, suggesting that quality and customer satisfaction are not solely determined by price. This directly highlights the products that are most favored by customers based on their ratings. This indicates strong customer demand and an opportunity to prioritize inventory and marketing efforts in this category.

Conclusion:

The analysis of the Amazon product dataset provided valuable insights into product categories, pricing strategies, and customer satisfaction. The data preprocessing steps were crucial for cleaning and transforming the raw data into a suitable format for analysis. Key findings indicate that customer satisfaction (ratings) is not directly proportional to price, and certain categories like Home Improvement and Electronics offer significant discounts. Products in Office Products and Toys Games, while lower in price, consistently receive high customer ratings. These insights can help businesses in optimizing pricing, marketing strategies, and product development to enhance customer satisfaction and sales.